The potential of *automated text analysis* for higher education research

by Stijn Daenekindt

Together with Jeroen Huisman, I recently published an article in which we mapped the field of research on higher education. In a previous blogpost we reflected on some key findings, but only briefly mentioned the method we used to analyze the abstracts of 16,928 research articles (which totals to over 2 million words). Obviously we did not read all these texts ourselves. Instead, we applied *automated text analysis*. In the current blogpost, I will discuss this method to highlight its potential for higher education research.

Automated text analysis holds tremendous potential for research into higher education. This because, higher education institutions—ie our research subjects— 'live' in a world that is dominated by the written word. Much of what happens in and around higher education institutions eventually gets documented. Indeed, higher education institutions produce an enormous amount and variety of texts, eg grant proposals, peer reviews and rejection letters, academic articles and books, course descriptions, mission statements, commission reports, evaluations of departments and universities, policy reports, etc. Obviously, higher education researchers are aware of the value of these documents and they have offered a lot of insightful case studies by closely reading such documents. However, for some types of research questions, analysing a small sample of texts just doesn't do the job. When we want to analyse huge amounts of text data, which are unfeasible for close reading by humans, automated text analysis can help us.

There are various forms of automated text analysis. One of the most popular techniques is topic modelling. This machine learning technique is able to automatically extract clusters of words (ie topics). A topic model analyses patterns of word co-occurrence in documents to reveal latent themes. Two basic principles underlie a topic model. The first is that *each document consists of a mixture of topics.* So, imagine that we have a topic model that differentiates two topics, then document A could consist of 20% topic 1 and 80% topic 2, while document B might consist of 50% topic 1 and 50% topic 2. The second principle of topic modelling is that *every topic is a mixture of words.* Imagine that we fit a topic model on every edition of a newspaper over the last ten years. A first possible topic could include words such as 'goal', 'score, 'match', 'competition' and 'injury'. A second topic, then, could include words such as 'stock', 'dow_jones, 'investment, 'stock_market' and 'wall_street'. The model can identify these clusters of words, because they often co-occur in texts. That is, it is far more likely that the word 'goal' co-occurs with the word 'match' in a document, then it is to co-occur with the word 'dow_jones'.

Topic models allow us to reveal the structure of large amounts of textual data by identifying topics. Topics are basically a set of words. More formally, topics are expressed as a set of word probabilities. To learn what the latent theme is about we can order all the words in decreasing probability. The two illustrative topics (see previous paragraph) clearly deal with the general themes 'sports' and 'financial investments'. In this way, what topic models do with texts actually closely resembles what exploratory factor analysis does with survey data, ie revealing latent dimensions that structure the data. But how is the model able to find interpretable topics? As David Blei explains, and this may help to get a more intuitive understanding of the

method, topic models trade off two goals: (a) the model tries to assign the words of each document to as few topics as possible, and (b) the model tries, in each topic, to assign high probability to as few words as possible. These goals are at odds. For example, if the model allocates all the words of one document to one single topic, then (b) becomes unrealistic. If, on the other hand, every topic consists of just a few words, then (a) becomes unrealistic. It is by trading off both goals that the topic model is able to find interpretable sets of tightly co-occurring words.

Topic models focus on the co-occurrence of words in texts. That is, they model the probability that a word co-occurs with another word anywhere in a document. To the model, it does not matter if 'score' and 'match' are used in the same sentence in a document or if one is used in the beginning of the document while the other one is used at the end. This puts topic modelling in the larger group of 'bag-of-words approaches', a group of methods that treat documents as …well … bags of words. Ignoring word order is a way to simplify and reduce the text, which yields various nice statistical properties. On the other hand, this approach may result in the loss of meaning. For example, the sentences 'I love teaching, but I hate grading papers' and 'I hate teaching, but I love grading papers' obviously have different meanings, but this is ignored by bag-of-words techniques.

So, while bag-of-word techniques are very useful to classify texts and to understand *what* the texts are about, the results will not tell us much about *how* topics are discussed. Other methods from the larger set of methods of automated text analysis are better equipped for this. For example, sentiment analysis allows one to analyze opinions, evaluations and emotions. Another method, word embedding, focusses on the context in which a word is embedded. More specifically, the method finds words that share similar contexts. By subsequently inspecting a words' nearest neighbors — ie which are the words often occurring in the neighborhood of our word of interest — we get an idea of what that word means in the text. These are just a few examples of the wide range of existing methods of automated text analysis and each of them has its pros and cons. Choosing between them ultimately comes down to finding the optimal match between a research question and a specific method.

More collections of electronic text are becoming available every day. These massive collections of texts present massive opportunities for research on higher education, but at the same time they present us with a problem: how can we analyze these? Methods of automated text analysis can help us to understand these large collections of documents. These techniques, however, do not replace humans and close reading. Rather, these methods are, as aptly phrased by Justin Grimmer and Brandon Stewart, 'best thought of as *amplifying* and *augmenting* careful reading and thoughtful analysis'. When using automated text analysis in this way, the opportunities are endless and I hope to see higher education researchers embrace these opportunities (more) in the future.

Stijn Daenekindt is a Postdoctoral Researcher at Ghent University (Department of Sociology). He has a background in sociology and in statistics and has published in various fields of research. Currently, he works at the Centre for Higher Education Governance Ghent. You can find an overview of his work at his Google Scholar page.

Available online on: https://srheblog.com/?p=1832&preview=1&_ppp=4a0854bc9e